Data Analysis on MovieLens

Xinwei Zhang

May 3, 2023

1 Introduction

As part of the Learning Data Science with Python course taught by Professor Junfeng Hu at Peking University in Spring 2023, I conducted a data analysis project on the MovieLens dataset.

The project involved a comprehensive analysis of the MovieLens dataset, including data cleaning, visualization, and modeling. The project's code is available on the GitHub repository, accessible at https://github.com/zachariah-zhang/Data˙Analysis˙MovieLen.

2 Task One

The first task of this analysis involves examining the movie preferences of different genders.

2.1

To measure a movie's popularity, the number of viewers is utilized as the primary metric in this analysis, which is a common and reasonable approach. To achieve this, all movies with a total rating greater than 300 are filtered and subsequently sorted according to the total number of viewers. The top 20 movies based on this metric are then selected.

The findings indicate that the top 20 popular movies for males and females differ to some extent. For males, the most popular movies are `['American Beauty (1999)', 'Star Wars: Episode IV - A New Hope (1977)', 'Star Wars: Episode V - The Empire Strikes Back (1980)'...]`, For females, the most popular movies are `['American Beauty (1999)', 'Shakespeare in Love (1998)', ...]`. The full list of the top 20 movies for males and females can be found in the code file.

2.2

The subsequent analysis focuses on examining gender-based movie preferences across various movie genres. Specifically, the number of viewers for each genre is calculated and compared between males and females. The results are visualized in Figure 1, which is included below.

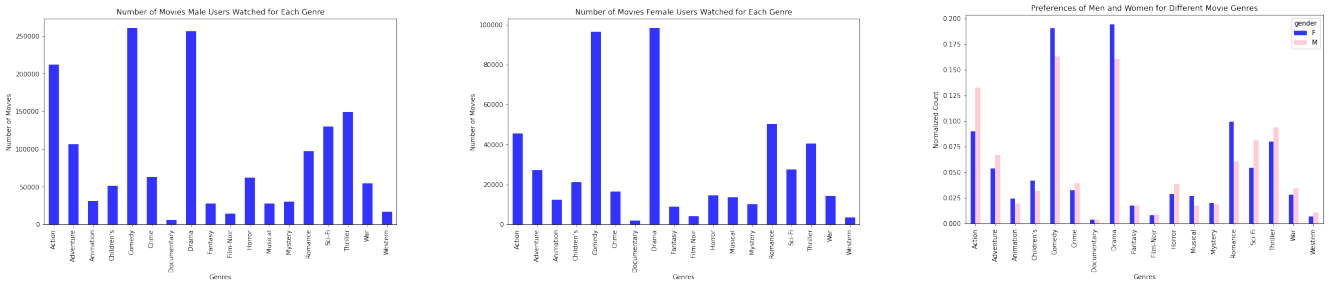

Figure 1: Visualization of Task One

3 Task Two

In the second task, a classification model is trained on the dataset to predict a user's gender and age interval. To leverage the entire dataset, the features of each user are defined as a vector consisting of two components. The first component is the average rating of the movies they have watched for each genre. The second component is the fraction of the total number of movies they have watched that belong to each genre.

By utilizing these features, a comprehensive representation of each user's movie preferences can be obtained, which enables the effective training of the classification model.

Two separate neural network classifiers are defined in this task to predict a user's age and age interval, respectively. Both classifiers share a similar structure, consisting of two linear layers with a relu activation function. Given the relatively small size of the dataset, the neural network models are trained using the entire dataset.

Figures 2, 3, and 4 illustrate the network structure and training loss of the AgeNet and GenderNet classifiers. Both models are evaluated on a test dataset after training, achieving accuracy scores of 0.8242 and 0.3956, respectively. The performance difference between the two classifiers suggests that predicting age is a more challenging task than gender classification. However, the high accuracy of the GenderNet classifier demonstrates the effectiveness of the feature vector used in this analysis.

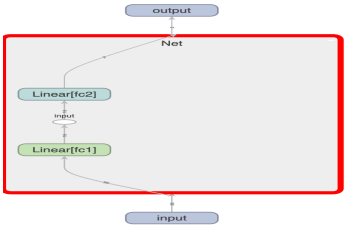

Figure 2: The network structure

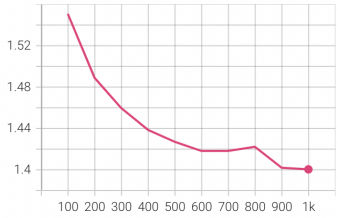

Figure 3: The training loss of AgeNet

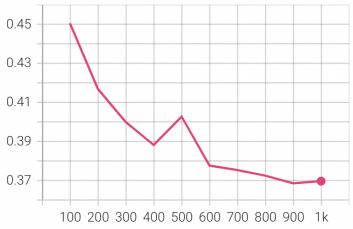

Figure 4: The training loss of GenderNet

4 Task Three

The final objective of this analysis is to generate a user profile for individuals who have watched more than 100 movies within the dataset. As previously established, the average movie rating and genre distribution are valuable features for constructing such a profile.

Consequently, the user profile can be created by plotting the distribution of the user's average movie ratings for each genre and the distribution of the proportion of movies they have watched belonging to each genre.

In addition, a user profile can be created for a specific user by identifying frequently occurring keywords in the movie descriptions of the movies they have watched. To achieve this, a word cloud can be generated to visualize the most commonly used words.

Figure 5 depicts the three visualizations created for user 2737, including the user's average rating distribution for each genre, the proportion of movies they have watched for each genre, and the word cloud of frequently occurring keywords in the movie descriptions.

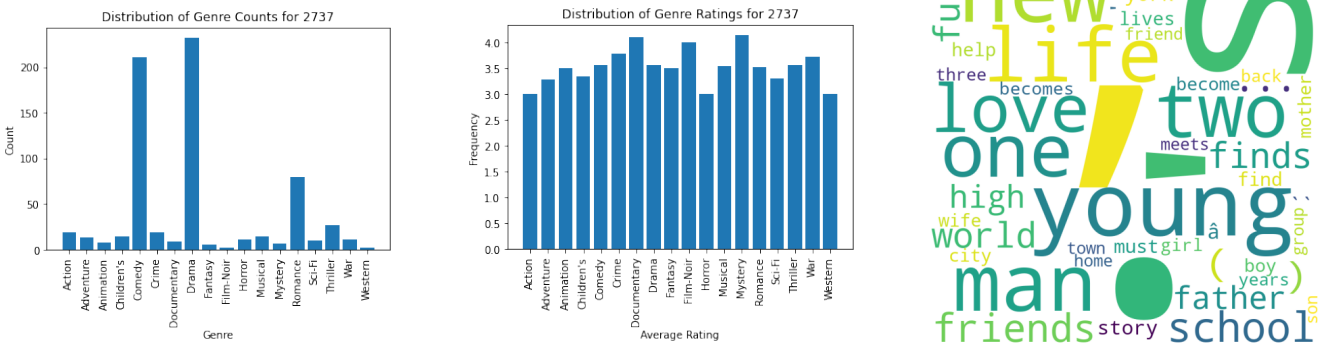

Figure 5: Visualization of Task Three